RCMOS – Multidisciplinary Scientific Journal of Knowledge.
ISSN: 2675-9128. Sao Paulo-SP.
Year V, v.1, n.1, Jan/July 2025. | submission: 2025-05-17 | accepted: 2025-05-19 | publication: 2025-05-21

Application of machine learning techniques to optimize hospital management

Application of machine learning techniques in hospital management optimization

Faitoma Jorge – Kimpa Vita University Nicolau Pedro – Kimpa Vita Nkanga University
Pedro – Kimpa Vita University

SUMMARY

This study aims to apply the K-Means clustering algorithm to identify disease patterns and support decision-making at the Catapa Hospital, through the integration of solutions based on emerging technologies, with emphasis on Artificial Intelligence and, in particular, Machine Learning, in order to improve the analysis of clinical data and optimize the management of health resources. The methodology adopted involved the collection and processing of 4,050 clinical records, including variables such as date of care, neighborhood, gender, age, age group, type of disease and neighborhood grouping. After preprocessing the data, the K-Means algorithm was applied, enabling the formation of 5 clusters composed of patients with similar characteristics. The analysis indicated that the Catapa neighborhood (classified as periurban) concentrates approximately 58% of the cases of malaria and fever. It was also observed that the age group of young people represents 38.4% of the records; children, 32.3%; adolescents, 14.9%; adults, 11.8%; and elderly, 2.6%, with a slight predominance of males (50.1%) over females (49.9%). The Catapa neighborhood alone accounts for 37% of the registered cases; Mbemba Ngango for 15%; Kindenuku for 10%; Dunga for 5%; Papelão for 3%; and the other 34 neighborhoods account for the remaining 30%. The month of June had the highest incidence, with approximately 36% of the total occurrences. It was also recorded that regarding the types of neighborhoods, the Periourban type had an occurrence of 73.6% of cases, urban neighborhoods with 24.6% and rural neighborhoods with 1.8%, according to data extracted from the hospital registry book.

Keywords: Machine Learning, K-Means, Hospital Management, Clustering, Data Analysis

ABSTRACT

This study aims to apply the K-Means clustering algorithm to identify disease patterns and support decision-making at Catapa Hospital. In this context, the integration of emerging technologies, particularly Artificial Intelligence and Machine Learning becomes essential for improving clinical data analysis and optimizing health resource administration. The methodology involved the collection and processing of 4,050 clinical records, covering variables such as date of attendance, neighborhood, gender, age, age group, type of disease, and neighborhood classification. After data preprocessing, the K-Means algorithm enabled the formation of five clusters composed of patients with similar characteristics. The analysis indicated that the Catapa neighborhood (classified as peri-urban) accounts for approximately 58% of malaria and fever cases. The youth age group represents 38.4% of the records, followed by children (32.3%), adolescents (14.9%), adults (11.8%), and the elderly (2.6%), with a slight predominance of the male gender (50.1%). Catapa alone accounts for 37% of recorded cases; Mbemba Ngango for 15%; Kindenuku for 10%; Dunga is 5%; Cardboard is 3%; and the remaining 34 neighborhoods for 30%. June had the highest incidence, with about 36% of all occurrences. Regarding neighborhood types, peri-urban areas reported 73.6% of cases, urban areas 24.6%, and rural areas 1.9%. Additionally, an interactive dashboard was developed using the Dash library, allowing dynamic visualization of data and results, providing hospital managers with a practical tool for evidence-based analysis and decision-making. The findings demonstrate that K-Means clustering is effective in identifying patterns and supporting hospitals

management.

Keywords: Machine Learning, K-Means, Hospital Management, Clustering, Data Analysis

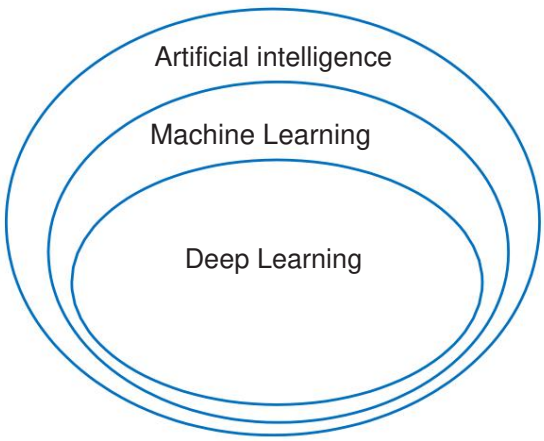
1. INTRODUCTION

Efficient hospital management is essential to ensure the quality of health services, especially in regions with limited resources, such as the province of Uíge. The General Hospital of Catapa (HGC), essential in serving the population, faces significant challenges due to the lack of automated disease monitoring systems. The absence of technological tools to predict outbreaks and carry out effective resource allocation compromises the ability to respond to disease outbreaks, worsening the health situation public. This context highlights the urgent need for an innovative solution that integrates advanced technologies, such as Machine Learning (ML), to improve hospital management.

The central objective of this work is to explore how the application of ML techniques can optimize the management of the Catapa General Hospital, helping to predict outbreaks and identify patterns of diseases that can improve resource allocation and preventive actions. Through analysis of collected health data, the aim is to identify the most affected neighborhoods, risk groups by age group and gender, and enable more accurate and faster decision-making. The using ML techniques, such as clustering, will allow not only the prediction of outbreaks of diseases, but also the targeting of strategies to minimize their impact and improve quality of service.

In short, this study aims to contribute to better management of health resources in Uíge. specifically at Catapa hospital, using artificial intelligence to predict diseases, improve control and optimize public health strategies in the province.

Figure 1: Relationship of AI, ML and DL



Source: Author

2. THEORETICAL FRAMEWORK

After carrying out a narrative review of the literature, we sought to present discussions conceptual issues on the technical topic of AI applied to hospital management processes. The “review narrative” does not apply explicit and systematic criteria in the search and critical analysis of literature. The selection of studies and interpretation of information may be subject to the subjectivity of the authors and this type of review does not need to exhaust the sources of information collection (Filho, 2020).

Machine learning encompasses the formulation of models or algorithms that acquire knowledge of historical data sets to facilitate predictions or take action. These models are trained using labeled or unlabeled datasets, and their effectiveness is enhanced as they are exposed to an increasing volume of data and constructive feedback (Pedro, 2024, p. 42). Understanding the principles, methodologies and tasks fundamentals inherent to this process constitute one of the fundamental pillars of learning machine. Techniques such as clustering, decision trees and random forest are included among the methodologies examined in this analysis.

“The fundamental principles of machine learning are examined in this chapter specific. It covers unsupervised methodologies (including K-prototypes and their associated validity metrics), as well as supervised approaches (such as machine learning support vectors, decision trees and random forests). We elucidate the mechanisms operational aspects of these techniques along with their corresponding performance metrics, which include precision, accuracy, recall, F1 score, and the confusion matrix. In addition, we investigate the SMOTE technique to highlight distinctions between datasets balanced and unbalanced”. (Pedro, 2024, p. 42).

2.1 Artificial Intelligence

Artificial Intelligence's main objective is to look for methods and ways of computers to do the same kind of analysis that the human mind does systematically. This definition of AI despite being quite simple to understand, the truth is that behind this such a brief statement does not reveal the great complexity and potential of this topic. By stating that

This concept relates to the human mind, we refer to psychological capacities as the

RCMOS – Multidisciplinary Scientific Journal of Knowledge.
ISSN: 2675-9128. Sao Paulo-SP.

prediction, association, perception, planning, among others (RUSSELL & NORVIG, 2016).

2.2 Considerations and justification for the choice of model

Hospital management is a complex field that involves the coordination of several resources, services and professionals with the aim of offering quality health care. With the exponential growth in the amount of data generated in the healthcare sector, the need for effective methods for analyzing and interpreting this information becomes increasingly evident. In this context, ML techniques emerge as powerful tools that can transform the way hospitals operate (SHAPIRO, 1992) This research adopted an advanced machine learning methodological approach to typify the index of diseases by different categories. The K-Means algorithm specified for categorical variables, enabling a more accurate and comprehensive representation of the characteristics of pathologies.

The choice of the K-Means algorithm was due to its effectiveness in dealing with structured data, its ability to form distinct groupings based on common characteristics and their efficiency computational. It was especially useful for identifying patterns in the distribution of diseases by region and age range, enabling efficient segmentation and supporting targeted strategies prevention and treatment.

3. MATERIAL AND METHOD

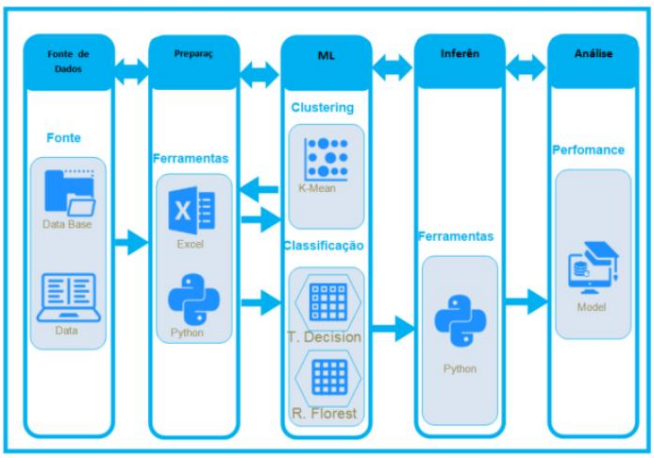
This research is characterized as quantitative, as it involves the collection, organization, analysis and interpretation of numerical and categorical data with the aim of understanding patterns and make predictions. The study also has an exploratory and descriptive nature, since seeks to identify behaviors, frequencies and groupings of diseases in different contexts (neighborhoods, age groups, genders, etc.) with the help of Machine Learning algorithms Learning.

3. 1 Project architecture

At this point, the main analyses of different methods and tools are presented and metrics used in the project. The study steps are illustrated in Figure 2

Figure 2: Project architecture

RCMOS – Multidisciplinary Scientific Journal of Knowledge.
ISSN: 2675-9128. Sao Paulo-SP.



3.1.1 Data description

3.1.1.1 Data source

To evaluate the different ML techniques used in this dissertation, a set of data relating to patients attending the Catapa Hospital, in the years 2022 to 2023. The dataset consists of 4050 records (patients with multiple records historical) with 08 attributes that were extracted from the HC record book.

3.1.1.2 Data preparation

After data collection, data was prepared in order to combine, structure and organize them to be used:

- To create machine learning models;

- Data analysis and visualization applications.

Using statistical tools and machine learning algorithms to discover the trend of patients registering at the Hospital.

3.1.1.2 Machine learning training

The machine learning algorithms suggested in this study, which allowed us to typify the diseases of patients by age group, gender and neighborhood, were trained using the language

Python and the SciKit-Learn library.

ÿ *Python:* It is a high-level programming language widely used for various several applications, including ML (Pedro, 2024)

ÿ *Scikit-Learn:* It is a Python module that integrates machine learning algorithms. state-of-the-art CA for supervised and unsupervised problems, focusing on ease of use and performance (Pedro, 2024)

3.1.1.3 Machine learning inference

According to Pedro (2024), he highlights that unsupervised machine learning models and supervised receive new data points during the inference process. To To begin, we use the K-Means model to group diseases according to neighborhood, gender, age, type of disease, type of neighborhood or similar characteristics. After that, the grouped variable, or cluster, is added to the new dataset to be fed into the training for the models. supervised machine learning that calculates the prediction of whether patients belong to a certain pathology as soon as they arrive. Model registration makes it easier to track models trained in the hospital.

3.2 Patient data structure

Information relevant to the disease prediction model includes elements that impact patient behavior. The following Table lists the data elements that affect patient attrition and retention patterns.

Table 1: HC Patient Data Structure

Feature	Type	Description
No.	Number	Registration number
Date	Categorical	Registration date
Patient Name	Categorical	Patient's full name
Neighborhood	Categorical	Neighborhood of origin
Gender	Categorical	Sex (Male and Female)
Age	Number	Patient age (0 to 150 years)
Onset of Disease	Categorical	Date of onset of symptoms
Complaint and Objective Examination	Categorical	Complaint of those who feel
Clinical Diagnosis	Categorical	Recording of the disease by the doctor
Results of Complementary Examinations	Categorical	The results are given
Indications	Categorical	Doctor's recommendations
Twelve	Number	Medication dose
No. of Days	Number	Days that should follow twelve
Quantities	Number	Amount of medicine (ml)
Weight	Number	Patient weight

Preprocessing was performed to "improve data quality through elimination or minimization of problems", including noise, incorrect values, inconsistent or absent.

Revista Ibero-Americana Multidisciplinary Scientific Journal of Knowledge.
ISSN: 2675-9128. Sao Paulo-SP.

Through data preprocessing, elements considered irrelevant were manually removed and which put the privacy of patient data at risk. The authors clearly state that when an attribute does not contribute to the estimation of the value of the target attribute, is considered irrelevant. In this case, of the 15 attributes in the dataset initial, 10 attributes were removed, and the other 5 attributes remained, which are Registration date, neighborhood, gender, age, clinical diagnosis (disease).

After exploratory analysis, the following attributes were added to each patient's situation, including age range, disease type and neighborhood type. This allowed us to analyze by disease and by range to find out which diseases are most frequent in the neighborhood of origin, establishing profiles and other characteristics.

We have 4050 records (patients) and 08 columns, the table shows the remaining attributes.

Table 2: Data Structure of the Processing

Feature	Description
Date	Registration date
Neighborhood	Neighborhood of origin
Gender	Sex (Male and Female)
Age	Patient age (0 to 150 years)
Clinical Diagnosis (Disease)	Registration of the disease by the doctor
Age Range	To group age by range
Type of Disease	Disease by type (Parasitic and Hypersensitivity)
Neighborhood Type	Grouping of neighborhoods by type (urban, peri-urban and rural)

K-means clustering algorithms were used as machine learning methods to solve the problem of disease prediction and clustering. The preprocessing techniques used included the treatment of unbalanced data, noise, incompleteness, redundancies and conversion of categorical data into numbers.

3.3 Patient Profiles Using K-Means Clustering

a) Contextualization

In the context of hospital management, the formation of patient profile groups based on clinical and demographic data is an essential step in identifying relevant patterns that can help predict outbreaks, optimize resources and define prevention strategies. intervention. The use of the K-Means algorithm allows grouping categorical data, enabling a more comprehensive analysis of the health scenario in the region studied.

In this study, variables such as: date of occurrence, neighborhood of residence, gender, age, age group, type of disease, neighborhood category and clinical status. The choice of these variables aim to capture different dimensions of the epidemiological profile of patients, contributing to the construction of a more efficient decision support system.

The main objective of this phase is to segment the population into groups with characteristics similar, in order to facilitate the identification of critical patterns of disease incidence, regional risk behaviors and possible vulnerable groups. From these clusters, the aim is to implement an early warning system and provide support for the formulation of more targeted and effective public policies in the health area.

b) Method for finding the optimal value of K

To determine the ideal number of clusters (k), an approach based on multiple iterations. The algorithm was run 20 times for each value of k, ranging from 2 to 40, and the average of the results was calculated in order to minimize the influence of values random.

Three validation methods were used: Elbow, Silhouette Score and Davies Bouldin Index.

Such a combination strengthens the robustness of the choice of k, considering both internal compactness of the clusters as to the separation between them.

The Elbow method indicated an inflection point at k = 5, at which point the rate of reduction of WCSS (Within-Cluster Sum of Squares) stabilizes. The Silhouette Score, in turn, presented the highest average also at k = 5, with variations between 0.0828 and 0.11 throughout the iterations. Although k = 4 has also demonstrated satisfactory performance, the choice of k = 5 is justified by the consistency of the results obtained between the different metrics. Subsequently, the K-Means algorithm was applied to segment the diseases into distinct neighborhoods, based on the optimal configuration of k = 5. The average silhouette score obtained confirms the quality of the segmentation, indicating that the generated clusters are well defined and clearly separated.

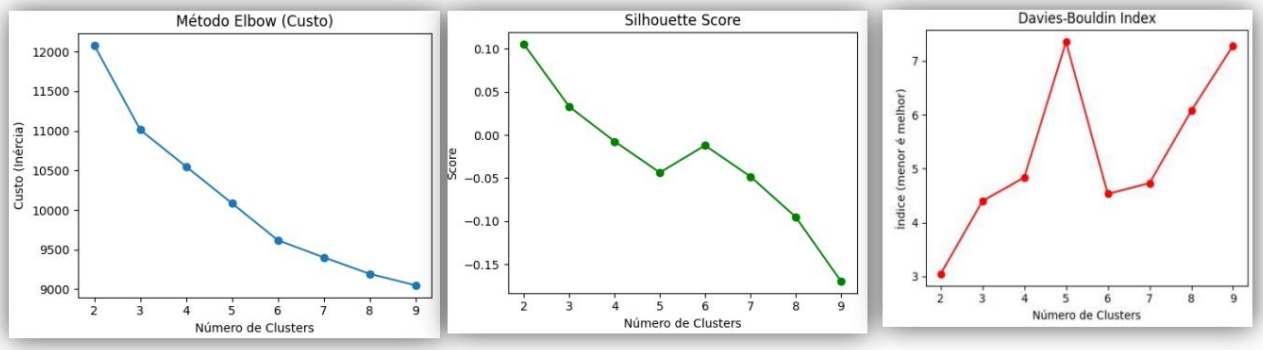


Figure 3: Validation Method

Note: The selection of 5 clusters was based on a joint analysis of three evaluation metrics: Elbow Method, Silhouette Score and Davies-Bouldin Index, taking into account not only the mathematical values, but also the practical interpretation of the generated groupings. 5 clusters were chosen as they presented a better balance between separability (good Silhouette), low cost (Elbow) and minimal overlap (smaller Davies-Bouldin), in addition to offer a clearer and more useful interpretation for hospital management.

the) ***View the profiles created after training the k-means model***

After defining the ideal number of k, using the Elbow, Silhouette Score and Davies-Bouldin, we applied the K-Means algorithm to segment disease cases into profiles distinct, considering both categorical and numerical variables of the patients treated at the Catapa Hospital.

As a result, five clusters (or profiles) were created, each representing a group of patients with similar characteristics, such as type of disease, neighborhood of origin, age group and gender. These profiles allow for a more in-depth look at the patterns of occurrence of diseases in different regions and population groups.

Figure 4 shows the distribution of cases by cluster, allowing you to see how the groups are formed. To reinforce our analysis and investigate the robustness of the segmentation, we chose to carry out two distinct approaches:

- *One including the "Age Group" variable;*
- *Another excluding this variable.*

This differentiation allowed us to evaluate how much the type of neighborhood influences the formation of groups of diseases, providing support for more solid hypotheses about the relationship between the environment and registered cases.

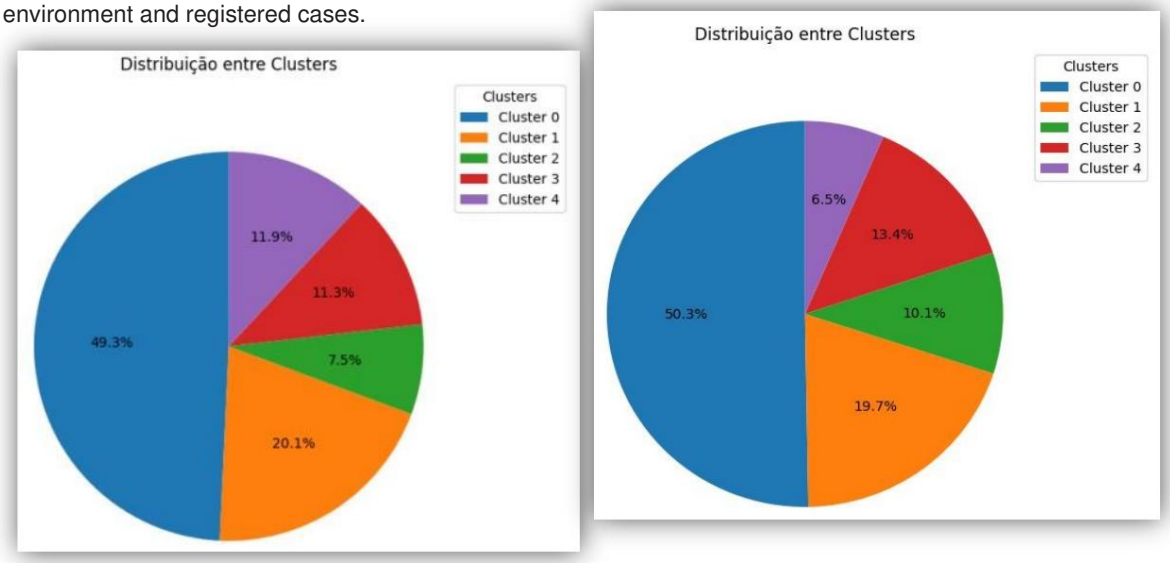


Figure 4: Distribution of Diseases in cluster with Age Group variable and without Age Group

After segmenting students with the K-Means model, we obtain the following results:

Cluster by Age Group

the)

- Cluster 0: 1995 Patients;
- Cluster 1: 813 Patients;
- Cluster 2: 303 Patients;
- Cluster 3: 456 Patients;
- Cluster 4: 483 Patients.

This organization allows a more detailed analysis of patient groups with similar characteristics. Each cluster represents a specific set of patients based on their characteristics. It is a useful approach to personalize teaching strategies and provide appropriate support to each group.

- In terms of distribution by Gender, Cluster 0: Predominantly male, with 80.5% men and 19.5% women, cluster 1: Here, men are the majority, with 69.3%, while women account for 30.7%, cluster 2: The distribution between genders is more balanced, with 51.7% men and 48.3% women, cluster 3: The group has a male predominance with 58% men and 42% women and cluster 4: The group is more balanced, with 55.2% men and 44.8% women.
- The distribution by Age Group, cluster 0: The group is composed mainly of young people between 19 and 29 years old (53.2%), followed by adults aged 30 to 39 (32.7%), cluster 1: A The predominant age group is 30 to 39 years old, with 61.5%, followed by young people aged 19 to 29 years old (24.1%), cluster 2: The majority of the group is in the 19 to 29 age range (63.4%), followed by adults aged 30 to 39 (36.6%), cluster 3: This group is predominantly composed of young adults, with 72.3% between 19 and 29 years old, and 27.7% between 30 and 39 years old, cluster 4: The largest group is in the age range of 30 to 39 years (68%), followed by 30.4% between 19 and 29 years.
- In terms of Diseases, cluster 0: Most cases are malaria (45%), bre (30%), and malaria (15%), cluster 1: Malaria predominates with 60%, followed by fever with 25%, and malaria with 10%, cluster 2: Fever is the most common disease with 50%, followed by malaria with 40%, cluster 3: Malaria has a prevalence of 55%, followed by malaria with 30%, and fever with 10% and cluster 4: malaria is the most prevalent disease (65%), followed by malaria (20%), and fever (10%).
- In Bairro, group 0: The Catapa neighborhood is the most affected with 58%, followed by Kindenuko with 20%, group 1: The Mbemba Ngango neighborhood has the highest incidence with 53%, followed by Dunga with 22%, group 2: The Kindenuko neighborhood leads with 62% of cases, followed by Kilala with 15%, group 3: Catapa is the neighborhood with the highest concentration of cases (50%), followed by Mbemba Ngango with 20% and finally in group 4: Catapa leads again with 60%, followed by Kilala with 25%.

b) *Cluster without the Age Group variable*

- Cluster 0: 2036 Patients;
- Cluster 1: 799 Patients;

RCMOS – Multidisciplinary Scientific Journal of Knowledge.
ISSN: 2675-9128. Sao Paulo-SP.

- Cluster 2: 411 Patients;
- Cluster 3: 542 Patients;
- Cluster 4: 262 Patients.

This organization allows a more detailed analysis of patient groups with similar characteristics. Each cluster represents a specific set of patients based on their characteristics. It is a useful approach to customize hospital strategies offer adequate support to each group.

• In the distribution by Gender, cluster 0: Male predominance with 80.5%, and the remaining female with 19.5%, cluster 1: The male distribution is 69.3%, with 30.7% of women, cluster 2: The gender is more balanced, with 51.7% men and 48.3% women, cluster 3: There is a male predominance of 58% men and 42% women, cluster 4: The group has a balanced distribution between genders, with 55.2% men and 44.8% of women.

• In terms of distribution by Diseases, cluster 0: Most cases are of malaria (45%), fever (30%) and malaria (15%), cluster 1: Malaria predominates with 60%, followed by fever with 25% and malaria with 10%., cluster 2: Fever predominates with 50%, followed by malaria with 40%, cluster 3: Malaria has a prevalence of 55%, followed by malaria with 30% and fever with 10%, cluster 4: Malaria predominates with 65%, followed by malaria with 20%, and fever with 10%.

• Distribution by Neighborhood, number 0: Catapa leads with 58% of cases, followed by Kindenuko with 20%, cluster 1: Mbemba Ngango is the neighborhood with the highest number of cases, with 53%, followed by Dunga with 22%, cluster 2: Kindenuko has 62% of the cases, followed by Kilala with 15%, cluster 3: Catapa again stands out with 50%, followed by Mbemba Ngango with 20%, cluster 4: Catapa has 60% of the cases, followed by Kilala with 25%.

3.4 Patient clusters with machine learning model

the) *Contextualization*

The methodology used in this study revealed a strong predictive capacity, as demonstrated by the accuracy of clustering models. The application of clustering techniques Machine Learning (ML) has been shown to be effective in analyzing healthcare data, enabling a deeper understanding of the patterns that exist among patients.

To predict the clinical status of patients, a grouping process was implemented, which enabled the development of differentiated profiles based on variables such as gender, location of residence, type of pathology and age group. These profiles received a new variable categorical called "Cluster", which summarizes similar characteristics between the groups.

By incorporating clustering as part of the feature engineering process, the model incorporated not only individual data, but also the collective context of patients, which had a significant impact on improving forecast accuracy. Thus, healthcare institutions can identify patients or regions at risk in advance or with greater chances of success in treatment, allowing the implementation of strategies specific to reduce losses, optimize resources and save lives.

b) Assessment instrument

We evaluated the performance of the decision tree and random forest models in predicting diseases through performance metrics such as precision, recall and F1-score. The findings demonstrated high accuracy, suggesting the effectiveness of the models created. The objective main thing is to identify the most appropriate forecasting model and compare the different metrics predictive performance of each group used. We started with unbalanced samples and subsequently balanced so that we can compare the results.

1) Prediction without sample balancing

Currently, we will examine the results of the prediction without sample balancing, observing how the models perform without adjustment in class proportions. We will see how this method can impact precision, recall, and other evaluation metrics. Let's examine this case and understand its implications.

4. RESULTS AND DISCUSSION

In this chapter, the main results obtained from the application of the K-Means algorithm. The analysis allowed to identify relevant patterns in the occurrence of diseases, considering variables such as neighborhood, age group and gender. The dataset included 4,050 records, covering patients treated between 2022 and 2023 at the General Hospital of Catapas.

The application of ML techniques in healthcare must consider ethical aspects, such as privacy of patients and equity in predictive models. Studies have shown that algorithms may exhibit racial bias and neighborhood influence, affecting equity in care. In addition, In addition, challenges such as data quality and resistance to adopting new technologies can impact the effectiveness of the proposed solutions.

Comparison between the ML techniques applied in our research and existing studies demonstrates that our approach is aligned with current practices in optimizing management hospital.

4.1 Analysis of Results

The analysis of the data obtained allowed us to observe significant patterns related to the incidence of diseases in different neighborhoods of the city of Uíge, considering variables such as age group, gender and types of neighborhoods. With the application of K-Means clustering algorithms, it was possible to segment the data into distinct groups, revealing specific concentrations of diseases in certain neighborhoods, which makes it easier for authorities to make decisions public health.

- *Analysis by Neighborhood*

Clustering revealed that peri-urban neighborhoods had a higher concentration of diseases parasitic diseases, especially among children and adolescents. For example, the Catapa neighborhood accounted for approximately 58% of cases of malaria and fever. The Catapa neighborhood, in isolation, accounts for 37% of registered cases, 15% for the Mbemba Ngango neighborhood, 10% for the neighborhoods Kindenuku, 5% of the Dunga neighborhood, 3% for the Papelão neighborhood and 30% of the cases is the total of percentages of the remaining 34 neighborhoods.

- *Analysis by Age Group*

The age group from 0 to 10 years old represents 32.3% of cases, the age group from 11 to 17 years of age account for 14.9% of cases, the age group from 18 to 45 years old accounted for about of 38.4% of registered cases, the age group from 46 to 59 represents 11.8% of cases and of the 60 years onwards correspond to 2.6% of cases. Most patients grouped in the clusters 0 and 1 were in this range, indicating high vulnerability.

Figure 5: Distribution of age group and type of disease

- *Analysis by Gender and Neighborhood Type*

In general, there was a predominance of males with a greater number of patients, around 50.1% of the records belong to men and 49.9% represent women. As for the type neighborhood, it was observed that the periurban neighborhoods predominate with an occurrence of 73.6%, urban with 24.6% and rural with 1.9%, as illustrated in figure 7

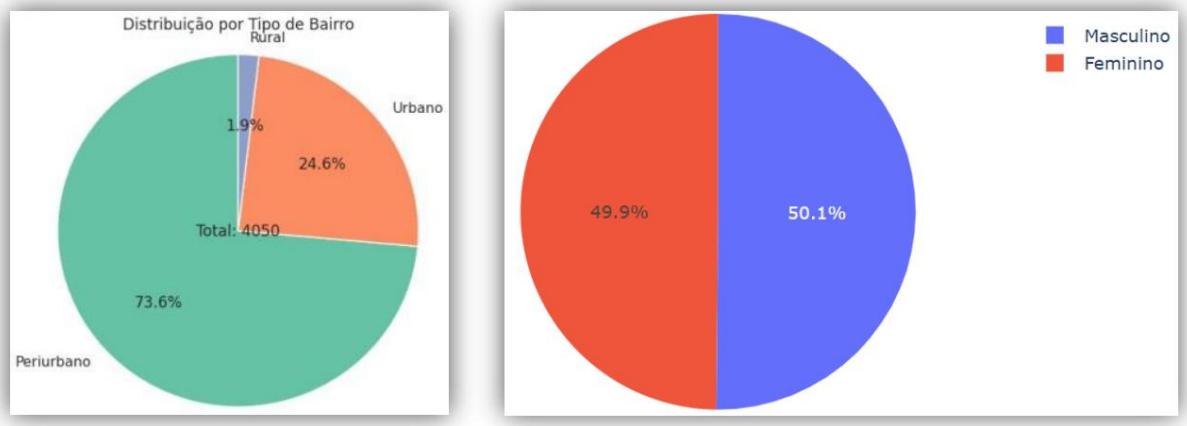


Figure 6: Grouping of neighborhood type and gender by range

- *Analysis by Month*

In all clusters, we observed that the predominant period is the month of June, highest incidence, with around 36% of total occurrences, according to data extracted from Catapa Hospital records book.

- *Identified Profiles*

Five distinct clusters were defined, representing groups of patients with similar profiles. Each group presents characteristic patterns regarding age, type of disease and location, contributing to more specific preventive actions.

- *Interactive Visualization*

The dashboard created with Dash allows you to view data interactively, offering filters by neighborhood, age group, gender and type of disease. This makes it easier for health professionals to analyze health and managers. In addition, the interactive dashboard developed with Dash proved to be effective in visualization of disease incidence and prediction patterns. The graphs and filters applied allowed data exploration in an intuitive way, supporting analyses for both experts and health managers with little technical familiarity. This panel contributes in a practical way to the transformation of data into useful and applicable information in context of hospital management.

FINAL CONSIDERATIONS

This work had as its main objective to apply Machine Learning techniques to optimize the analysis and data management at the Catapa Hospital in the municipality of Uíge. Through the implementation of clustering and grouping algorithms, it has been demonstrated that it is possible identify relevant patterns in disease distribution, anticipate potential outbreaks and support decision-making by health authorities.

Despite the limitations related to manual data collection, the scarcity of records digitized and the absence of a hospital management system, the results obtained were significant. The techniques applied showed that even with data obtained in environments with limited infrastructure, it is possible to extract value and generate knowledge that can be used to improve the allocation of resources and prevention actions.

The study demonstrated that the use of the K-Means algorithm is effective for segmenting populations with based on clinical and demographic data. The analysis revealed relevant patterns in the distribution of diseases, offering support for decision-making in the hospital context. The use of K-algorithm Means, provided different perspectives on the data, allowing not only segmentation them, but also predict occurrences with good accuracy. Creating an interactive dashboard

contributed to making these analyses accessible and applicable in the daily lives of institutions health.

Removing additional algorithms allowed for a clearer focus on clustering, without compromising analytical quality. The interactive dashboard expanded access to information, promoting data-driven management.

It is recommended, in future studies, to explore other algorithms with new data sets and compare the results with those obtained by K-Means. In addition, the expansion of the database data to include later years can enrich the analyses. As future work, it is also recommended to expand the database, integration with health systems in real time, and the inclusion of environmental variables, such as climate and sanitation, to increase predictive capacity of the models. The continuation of this work could contribute significantly to the advancement of data-driven public health in the country.

REFERENCES

Filho, CR (2020). ARTIFICIAL INTELLIGENCE TECHNIQUES APPLIED TO HOSPITAL MANAGEMENT PROCESSES. 11.

Marius Sumanas, AP (May 21, 2022). Deep Q-Learning in Robotics: Improvement of Accuracy. p. 16.

Maronna, R. A. (2017). Norman Matloff (2017): statistical regression and classification: from linear models for machine learning. *classification: from linear models to machine* learning(National University of La Plata).

Mechelli, A. (2019). ML Methods and Applications. *Machine Learning.*

Mitchell, T. (1997). Machine Learning. 1(Machine Learning is the study of computer algorithms that improve automatically through experience.), 9.

Mohri, M. (2018). Foundations of Machine Learning. *second edition,* 505.

Neves, S.A. (2018). Machine Learning Techniques . *TFC,* I(Quality Classification of Asphalt Pavements), 49.

Pedro, N. (2024). Use of machine learning techniques in an academic context for typification of the risk of school dropout. *Thesis_MSC_Final_Rev,* 91.

RUSSELL, S., & NORVIG. (2016). Artificial Intelligence. *Pearson.*

Samuel, AM (2011). *Methodological proposal for solving equations reducible to the second degree IR. Case of biquadratic and irrational equations.* Isced, Uíge, Agola.
SHAPIRO, SC (1992). Encyclopedia of artificial intelligence.

Sutton, R. S. (2018). Reinforcement Learning. *An Introduction, Second* edition (Complete Draft), 38.
Zhang, W. (2024). International Journal of Mental Health Promotion. p. 26.

.

ACKNOWLEDGMENTS

We would like to thank the Teaching and Research Department of the Institute Polytechnic of Kimpa Vita University in Uíge, which provided us with the necessary conditions for the completion of this study

To the Catapa Hospital, we express our sincere gratitude for the collaboration and support provided during the development of this work. The openness to access data and the welcome from the team was fundamental for carrying out the research and for the enrichment of the results presented here.

To **Sebastião Afonso,** we are deeply grateful for the companionship, for the exchange of knowledge, for the friendship built throughout this stage, for the intellectual support, for the guidance and moral support in the most challenging moments.